Comparison between the two definitions of AI*

Dimiter Dobrev
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
1113 Sofia, BULGARIA
e-mail: d@dobrev.com

January 17, 2013

Abstract

Two different definitions of the Artificial Intelligence concept have been proposed in papers [1] and [2]. The first definition is informal because it says that the programs that are cleverer than human are acknowledge as Artificial Intelligence. The second definition is formal because it avoids reference to the concept of 'human'. The readers of papers [1] and [2] are left with the impression that both definitions are equivalent and the definition in [2] is simply a formal version of that in [1]. This paper will compare both definitions of Artificial Intelligence and will hopefully help for the understanding of the concept.

What is the main idea?

The idea behind the definitions of Artificial Intelligence in [1, 2] is as follows. If a program is intelligent it would manage well in an arbitrary world. This is a version of the popular wisdom that the clever person can handle any job. Certainly, the clever person will manage not immediately but after some training (learning).

Therefore, when a program is measured how it manages in a particular world, it should go through a certain period of training. Only when the training is over, we should assess how well the program is doing.

We can make an analogy with humans. The first eighteen years of human life are considered as a period of training, if we do wrong and even commit

*This work was supported by Bulgarian Ministry of Education, project DID02-28.

a crime, the punishment will be weaker, because it is believed that we are still on training. The training period with animals is usually shorter; this is usually associated with their shorter lives and the worse living conditions. The rule that is often true about animals is 'Learn fast or be eaten!'.

How long is the training period and is it possible to say when exactly it is over?

Here comes the first difference between the definitions in [1] and [2]. The first definition supposes that life is infinite and there is enough time for training. This is to say, the training period is an arbitrary finite beginning of life, which is infinitely less than the whole life that is infinite.

The approach of the second definition is different. It assumes that life is limited and there is a parameter will be referred to as 'the maximum length of life'. The assessment of intelligence is based on life – from its beginning to the end. The second definition is different with view of the training period, as well. There is no period of training with the second definition. The rule which is true is 'Learn fast or be eaten'.

Why the period of training is zero with the second definition? Because we cannot say how long should be this period and when it will be completed. Therefore, it is convenient to assume that there is no such period.

Are both definitions equivalent?

It follows from the above that both definitions are not equivalent because the retarded program is Artificial Intelligence according to the first definition, however this is not true according to the second definition. Retarded is a program that needs nearly infinite time for training (i.e. its period of training is finite but in practice it is infinite). Examples for retarded programs are TD1 and TD2 in paper [2].

The retarded program does not match our idea of Artificial Intelligence. Our idea is to construct a robot that can sweep the floor and if the robot needs 1000 years to learn how to sweep, this would not work for us.

Are there other programs that formally satisfy the definition but do not match the idea of Artificial Intelligence? Yes, the extremely inefficient program is such a program, even if it satisfies both definitions. Paper [2] describes such a program as Trivial Decision 5. This program would work if we had an infinitely fast computer, as it makes nearly infinite number of steps (i.e. finite many but in practice infinite) in order to calculate only one step.

Both definitions define a set of programs. These sets should coincide in order to make the definitions equivalent. Therefore, both definitions are not equivalent and the example for this is the retarded program.

There are two additional reasons why both definitions are not equivalent:

Firstly, the definition in [1] is informal and dependent on people. That is to say, it does not define a particular set but rather a sort of fuzzy one. We have said that we acknowledge for Artificial Intelligence these programs that are cleverer than human. The reasonable question to ask with that statement is: 'Who is the man we compare with?' A possible answer is that the program is cleverer than any human, however, this would not define the set of these programs in a unique manner. For example, if we define the chess-playing program like a program playing chess better than the world champion, there will be still programs that will play chess better than one particular world champion and worse than another. Therefore, a formal and an informal definition cannot be equivalent.

Secondly, the definition in [2] is dependent on several parameters. This is to say, that this is not the case of set of programs but rather a function that for different parameter values returns different sets of programs.

The assumption made in [2] is that such parameter values exist for which the set defined by the definition is not empty and the programs within that set match our idea of Artificial Intellect.

Certainly, not all programs within that set match our idea. We have already mentioned the problem with the infinite ineffective program. Another problem that appears with the definition in [2] is the 'cramming' program. The problem is that in [2] it is assumed that the worlds we are interested in are finite in number. Therefore, a program can be created specially for these worlds. We come across the same problem with the candidate students – when people who have crammed all possible exam topics sit for the exam. Such people will manage and will pass the exam but they cannot solve any problem out of the range of the crammed exam topics.

The assumption is that if the shortest and the most efficient program is taken out of the those that satisfy the definition, then it will match our idea of Artificial Intelligence. We should limit the programs both in length and in efficiency because the infinitely inefficient program is rather short, whereas the 'cramming' programs are rather efficient. The 'cramming' program is shorter than the Artificial Intellect for a small number of worlds; however the Artificial Intelligence is a shorter program than the 'cramming' one if the number of worlds is sufficiently great.

Lifetime

The first parameter which the second definition depends on is the lifetime. Once we give up infinite life, then we should limit it and set a parameter

indicating the life expectancy. To make it simple, the life in [2] has been fixed at 100 games, each one no longer than 1000 steps.

Giving up infinite life, we get rid of the retarded program. Another advantage is that it is no longer required that the worlds are without fatal errors. This requirement was important in [1] when we needed sufficient time for training, presuming that no fatal error is made that would ruin the whole life. This is to say, we can easily make mistakes as each one can be overcome and none is fatal.

Definition: Fatal error is a group of internal states of the world, such that when we enter this group we can no longer exit. If there is an exit, the error would not be fatal. Besides, the world in this group should be worse than that out of the group (i.e. the assessments made there are relatively lower). If the world in the group is not worse, then the fact that we happen to be there would not be a mistake.

Another possible definition of fatal error is the following. If we calculate for each moment of life what is the maximum anticipated success of life (what returns the Success function), provided we play (live) by the best possible strategy, from this moment onwards, then fatal error will be referred as a step after which this number is decreasing.

Having assumed that life is finite, it is not needed to assume that there are no fatal errors in the world, because our time for training is anyway limited. When life is finite, a common mistake can be equal to fatal, because time may not be sufficient to fix it.

Thus, it is natural to assume that life is finite and we are looking for a program that can manage well within certain lifetime and not with any lifetime. On the other hand, it is inconvenient that there are parameters included in our definition. It would be better to define the Artificial Intellect as a program independent of anything. That is one program, irrespectively of the anticipated lifetime.

However, life expectancy is an important parameter that influences the strategies. Let's consider the human behavior in war time, natural disasters and other catastrophes. When life expectancy becomes shorter, the behavior of people significantly changes. This is expressed mainly in the tendency to take greater risks. You may also notice that the young people are braver than the older. A possible explanation of this observation is that the young are more willing to experiment and take risks while the adults prefer the stable and the secure because they estimate that there is no time for experimenting. Thus, we can say that life expectancy definitely influences behavior of people and their life strategies.

Arbitrary world

The first definition requires for the Artificial Intellect to manage as good as human in an arbitrary world. This requirement is so strong that it appears that there is no program that can meet the requirement and the set of programs satisfying the definition may prove to be empty.

Let's try to create a world which is too complex for any program to manage but not for the human. Imagine a world where robots are not liked. In such world, if you are recognized as a robot you immediately score low. However, if you are considered human you score high. It seems that this is the world where humans will do better than robots. Let's remember the definition of a world; there are two functions (World and View) defining the world. They are absolutely arbitrary functions and we can presume that they return success when there is human living in the world and respectively lack of success in the case of a robot. Still, the world is not God and there is no way to know if its inhabitant is human or robot. The world will know who is who based on the acts of the opponent. That is, if the robot behaves as human and acts appropriately then the world will be deceived and will accept it as human. In this case, the program is required to play the imitation game. The same game was proposed by Turing as test for Intelligence. It is to be concluded that if a program meets the definition in [1], then it is satisfactory to the Turing test, not immediately but after some time of training.

Question: Is it possible that the world recognizes the robot (while it is still on training and has not started to act like a man) and starts scoring low from that first moment of recognition till infinity? The answer is: No, because only worlds without fatal errors are considered and this world does not meet this requirement.

Does that mean that the definition in [1] is equivalent to the Turing test? Not if we train the program in [1] to pretend being a human, then it will satisfy the Turing test but only after training. Is it possible that the program satisfying the Turing test to be trained to manage in an arbitrary world? The answer is: rather not. If the program can pretend to be human, then it can be trained. However, it should rather pretend to be stupid and hide its intelligence otherwise, it will betray the fact that it is a robot and not human. If it is forbidden for the examiner to punish the excessive intelligence, then the definition in [1] and the Turing test will be equivalent.

Impossible World

Is it possible that the world is so complex that there is no program that can understand it? Yes, it is. For example, let the world generate an infinite row of zeros and ones. The Artificial Intelligence has been given a task to have a guess what comes next (zero or one). Let the function describing this infinite row is not computable. Then, there is no way for the program to calculate and make a guess which number will follow. This is true about the human, as well. Nevertheless, the program and the human will find different dependencies. For instance, such that the zeros are more than the instances of having one, that it is more likely that one comes after zero than zero and etc.

It is not necessary that the Artificial Intelligence should understand the world at 100 %. What is important is to understand the world better than the human.

IQ (Intelligence Quotient)

The first definition has compared the intelligence of the program with that of human. The second definition cannot allow making the comparison with man (because we want the definition to be formal). Therefore, it is necessary to introduce an independent assessment of IQ by which we could define the Artificial Intelligence. We will say that for AI we acknowledge those programs whose IQ is above certain value. This value was decided to be 0.7 in [2] but this choice has been largely arbitrary. It is rather correct to say that certain IQ exists and the programs more intelligent than this level are acknowledged as Artificial Intelligence.

We introduce the function Success which returns a number in the interval [0, 1] for each particular life. This number makes assessment of the device success in the particular life. Afterwards, the IQ is calculated selecting a set of test worlds, running the program to live a life in each one of these worlds and calculating the average success of the program in all its tested lives.

Thus, the IQ is the average value of Success function calculated based on the set of tested worlds.

World Complexity

Another substantial difference between both definitions is that the first considers all possible worlds; whereas the second limits the sets of the worlds to a finite number of tested worlds (the assumption is that the tested world is

computable with fixed level of complexity). This fixed level of complexity is the next parameter of the definition.

Why did we select the set of the tested worlds to be what it is?

Something similar has been done in paper [1]. It has proposed to prepare a test consisting of finite or countable number of tested worlds. The idea is to acknowledge as Artificial Intelligence the program that can manage in all these worlds. Paper [1] has proposed that these worlds are prepared by human but we want to be maximum formal in paper [2] and therefore we will define the set of tested worlds in such a way that it would not depend on human's choice. The other difference is that in [1] we want the program to pass all exams, i.e. to manage in all tested worlds, whereas in [2] we want the average success (i.e. IQ) to be greater than 0.7. Why we want to have less in [2] than in [1]? Because if the problems are preliminarily prepared we would want that the program will solve them all, but if the problems are randomly generated then there will be such that are not solvable and therefore we cannot rely that the program will do them all.

What will be the set of worlds that we will use for the calculation of the IQ of an arbitrary program?

The first natural possibility is to take the set of all worlds. This set is infinite, even uncountable and seems too large (it is not clear what should be the weight that different worlds will participate with). The first thing we find is that many of the worlds are indistinguishable (i.e. their tree of the world is equal). That's why we resort to the next idea that is to take the quotient set i.e. the set of all possible trees of the world and make it our set of tested worlds. This set is again uncountable, but considering the fact that we limited the lifetime, we see that the set of these trees is even finite (more precisely, the set of the trees of determined worlds is finite. It is finite with the undetermined worlds, as well, because the branches are equally probable to happen – see the definition of TM_W in [2]).

This set is not suitable (although it is finite) because anything is possible in such a case! How will the world of the next step continue to be? It may continue as it likes to. Anything is possible, indeed, but far from anything is likely to happen. If we accept this set as a tested one, then any continuation will be equally probable and what has happened so far will be of no importance. This totally contradicts our idea of Artificial Intelligence which says that the device gathers experience and is on training. Thus, what has happened so far is important.

This is the time to apply the principle known as 'The Occam's razor' stating that the simpler model is more probable than the complex one. Therefore, the simpler world is more probable than the complex one. If we are to discuss the complexity of the world, then we would introduce a description of the

world and define the complexity of the world as the length of the shortest possible description.

We have used the Turing's machines in [2] to describe the worlds. This is not the most suitable model in case you try to make a real program which satisfies the definition. However, it works just as a theoretical model of computability. Still, we do not want to be limited within the set of the determined worlds and this is why we have introduced undetermined Turing's machines. Thus, our tested worlds are the computable worlds generated by the undetermined Turing's machines.

The next suggestion is to select the set of tested worlds to be the set of the undetermined Turing's machines taking on board all such machines regardless of their size. Could we chose a particular size and do only with these machines? The answer is: rather, yes.

If we take all Turing's machines, then we should give them different weights. As there is no way that infinitely many machines are of equal weight and the sum of their weights to be one. Having decided what will be the weights of the different machines, there are two options to go for: either the average size (length) of the machines is a particular number, or the average size is infinity (depending on the weights we have finally chosen). If the average size is finite, then we can assume that instead of having all Turing's machines, only those with average size will be the tested ones. This is not same but is almost the same. If the average size is infinity, then there is an N and all machines with greater size would almost make no influence. Therefore, is we chose an ε that seems to us small enough to be ignored, then such N exists that the machines longer than N will influence less than ε of the average success. Then we can decide that N is the size of the tested machines and the result will be closer to that which would get at, if we consider all machines with their respective weights.

Next question: If we have decided on a particular N, should our tested machines be these of size smaller or equal to N or those with the size of N exactly. The answer is: there is no need to include the shorter machines because each machine with size $N-1$ has many equivalent machines with size N (because we can add a state that is not necessary).

So far, so good. We have decided that the tested worlds will be the computable worlds that are computable by an undetermined Turing's machine having the size of N. This makes the next parameter that our definition depends on. We decided on a particular value of 20 for this parameter in [2]. We decide that all tested machines will participate with equal weights (this is possible because the set is finite).

Does the set chosen in this way correspond to the principle of Occam? Are the simpler worlds more probable than the complex ones? The answer

is: yes. Indeed, all machines participate with equal weights but the simpler machines have a great number of equivalent machines (such that compute the same world), whereas there is not a single equivalent machine for the most complex ones (certainly, of the same complexity, in this case with complexity value of 20). This is to say that the simpler the world is, the more machines with size 20 can compute it and the more this world would influence the average value of Success function. We refer to this average value as IQ.

Which is the suitable model?

We have already said that the Turing's machines are not the suitable model to describe the world. We would like to have simple dependencies in the world that are on the surface and easy to be discovered, whereas more and more complex dependencies to be discovered the deeper we go. The Turing's machine is a dependence that may appear to be rather complex but once you understand it, you have understood the world. The case of the undetermined machines is better because the randomness is an infinitely complex dependence. Therefore, we will never understand this dependence because once we understand it; it will become pseudorandomness (take the example of pseudorandom numbers generated by the computer).

Is it possible that a complex Turing's machine is partially described by means of simpler dependencies? This is possible but it is not typical for the Turing's machine model. In this model usually, you either understand the whole world or you do not understand anything.

If we are looking for a world model of the type of determined machine, then very soon (i.e. after very short life experience) it would turn out that the first Turing's machine corresponding to that life experience is so complex that virtually it is impossible to find it. The advantage of the undetermined machines is that we will always be able to find such world model (no matter how long is the life experience). It is another point how adequate is this model and how good it will work for us, because the undetermined machine does not say what will happen with the next step but it rather says that this or that can happen. In the best case, it provides the probability of having this or that happening.

Well, which is the suitable model of the world? We should think of the world as a union of different factors that may be related but are largely independent. Certainly, we will need a better model if we want to create a particular program satisfying the definition of Artificial Intellect, but this is not important for definition itself.

Work of other researchers

The occasion to write this paper is publication [5] where two scientists from Switzerland have tried to generalize the definitions in [1, 2]. Their idea was to get rid of the parameters, which the definition in [2] depends on and to get to a new concept of IQ independent of any parameters.

They have removed the limitation on lifetime and have assumed that life is infinite. In their representation the beginning of life is the most important part of life. They assume that the rewards become lighter at any next step having been multiplied by the coefficient of discount.

Surely, the coefficient of discount is a parameter, as well, and they have simply replaced one parameter with another. What is more, their presentation is contrary to the idea that the beginning of life is not important. What is important is what happens having the program already been trained. Their presentation assumes the beginning of life as the most important part. The life is so much discounted that from a moment onwards, in practice, it is not important what the program is doing. Certainly, there is a moment (the maximum length of life) in [2] from which onwards there is no point what the program is doing, but at least until that moment all rewards are of equal importance. This is to say, that 'it runs, runs and stops' in our case, whereas 'it fades, fades, fades and like this to infinity' in their case.

We have to acknowledge that the authors of [5] have understood that the coefficient of discount is a parameter which their definition depends on and have, therefore, proposed a second version. It would have been better if the second version have not been proposed, at all, as it has resulted in meaningless outcome compromising the whole paper. More, about the second version, has been written further below (striking mistakes 3 and 4).

The other parameter, that the authors of [5] have tried to get rid of, is the world complexity. We have decided on particular complexity (i.e. number of states of the Turing's machine which generates the world). They have preferred to sum up all complexities having used a coefficient of discount 1/2. This results in having average complexity of 2 in their case (this parameter has been picked to be 20 in our case). This is to say, that if they want to allow for higher values of their average complexity, they will have to replace the number 1/2 with a different parameter. Therefore, they replace one parameter with another, again.

They get rid of the parameter 0.7 by omitting to say what Artificial Intelligence is. They say what is IQ but they do not say how big the quotient should be in order to acknowledge a program as Artificial Intelligence.

Unfortunately, our Swiss colleagues have omitted to quote the Bulgarian source. Another problem is the fact that they have not managed to un-

derstand many details of the original papers and, therefore, there are many mistakes and wrong things in the resultant text.

Striking mistakes

These are six of the most striking mistakes made in [5].

1. Their paper has written that the world is computable according to the thesis of Church. It is true that it has been written in [1] that it follows from the thesis of Church that the Artificial Intellect is a program but not that the world is a program, too. Is the world computable or not, is it determined or not, these are questions whose answers we do not know and will never know. This is something that cannot be verified because there is no such experiment that can result in the answer of these questions. (The question is the world determined has been considered in details in [3]. The question is the world computable is analogical).

2. They have defined the IQ in such a way that it is infinity for each program. It was not envisaged that the number of programs grows exponentially with the increase of their length. This can be regarded as an oversight mistake, moreover that it is clear how it can be fixed. (in their case, they sum up the programs' success achieved in different lives, i.e. the values of function Success. This sum does not require that each addend is multiplied by 1/2 raised to the power of the complexity degree, but would rather take the average for the respective complexity and multiply it by the same). It has been said above that the average complexity of the world in [5] is 2, having assumed that this mistake has been fixed. If what has been written remains unchanged, then the average complexity is infinity and the sum that we refer to as IQ is infinity, too. Thus, if the mistake is not fixed, the IQ concept becomes meaningless.

3. The most serious problem in [5] is the second version that was proposed in order to avoid the coefficient of discount. The set of worlds, in this version, is different and respectively the Success function is different (this function that says for each life what success has been achieved in this life).

Let's summaries the outcome in [1], [2] and both versions in [5].

An infinite life in a set of worlds without fatal errors is considered in [1]. It is assumed that life is finite in [2]. The first version proposed in [5] describes life as infinite but fading which is the same, as if the life is finite. The second version proposed in [5] assumes infinite life in a world where all errors are fatal. The problem, in this case, is not that there could be a fatal error (the fatal errors do not interfere in [2] and will not interfere in this case, as well). The problem is that all errors are fatal. This is to say, that there

are not fixable errors. Humans do not learn from their fatal errors, perhaps, they have learned from the fatal errors of others but not from their own. That's why this concept contradicts the idea of learning.

The Success function is monotonically increasing in the second version proposed in [5]. It was set to be the sum of all rewards which are numbers in the interval [0, 1]. It is natural to consider that the Success function can be increasing and decreasing during the lifetime. It can be considered as a fixable mistake, when it starts to decrease. The authors of [5] have preferred to have this function monotonically increasing which means that the device cannot make fixable errors. The only possible mistake is of the type 'lost profits' and this mistake is always fatal because once the profits are lost there is no way to bring them back. Besides, there is no feedback with the lost profits. Thus, when the Success function changes, the device will know but there is no way to know when the profits are lost. Eventually, it may know in future, but life is infinite and therefore, it will not know even in future. The device will always hope that the profits are not lost and will soon emerge.

4. The decision of the authors of [5] to limit the sum of the rewards is very strange and illogical. It reminds me of one of my teachers, for who the students were telling that he had a limit on the A grades and you should be among the first to be examined because the A grades will end. No matter how much knowledge you demonstrate if you are at the end of the exam you will not be A-graded because of his limitation.

This limitation is imposed, in their case, so that the Success function will be in the interval [0, 1]. Instead of distorting the world in such a horrible way, it would be better if the Success function is the arithmetical mean of the rewards (as it was done in [2]) instead of being the sum of the rewards. This would have made the Success function being in the interval [0, 1], and it would not be monotonically increasing.

The conclusion is that the authors of [5] use worlds where training is impossible in their second version of the IQ definition. The success of the device in such a world depends solely on being lucky, however when there are many worlds the luck ceases to act. Thus, all programs are equally intelligent.

The question to ask is whether the authors of [5] have managed to understand that both [1] and [2] consider a device that is being trained and will achieve good success as a result of the training, or they rather believe that the device was born trained. As a matter of fact, the beginning of [5] says that the device should be given sufficient time for training, however, later they propose two versions of the definition which contradict this idea (especially the second version).

It is true that with the definition of Turing, there is a device that was

born trained, but this concerns a particular world. The device can be born trained for a particular world but there is no way that it was born trained for any world.

5. We can consider for a mistake the fact that the definition of a world in [5] has been changed. The world in [1] has got a set of internal states and a function indicating how to transfer from one state to another. As you know, there is a tree of the world corresponding to each world. It was the tree of the world that has been regarded for the definition of the world in [5]. This is the same as if we do not consider functions in mathematics but only graphs of functions. We can somewhat justify the authors of [5] because they make their attempt to improve somehow the definition of AI, however their change implies that they have not understood our main idea. We suppose that the world has got some structure and the device tries to understand that structure. They deny the structure of the world and this is a mistake.

They act analogically when they define the device, as well. The device is a program for us, whereas for them it is a strategy. Certainly, there is a strategy corresponding to each program (but not vice versa). Still, to consider the device as a strategy is a mistake, as thus we presume that it has not got internal states, i.e. no memory. This mistake is very common among researchers working in the area of AI. Many of them look for the AI in the set of the functions meaning, that for them, AI is a device without memory. The strategy is also a function, whose input argument is the entire life experience. At first sight, it looks as if we do not need the memory if we have available the entire life experience, but this is not true.

Imagine, that at some point (based of your whole life experience) you decide to go to the fridge and grab a beer. Ten seconds later, you see that you are on your way to the fridge but you do not know if you are going to get a beer or milk. It is true that you can rely on your whole life experience but this will not answer the question. If you were told to fetch a beer, this can be extracted from your life experience, but if you have, yourself, decided to grab a beer, you would not remember it (because you have no memory) and there is no way that you extract it out of the life experience. This is to say, the memory is needed. We have to note that it is absurd that the device will make a decision at any step based on its whole life experience (because this is a huge amount of information). It is more logical to assume that it decides based on its internal state and the immediate input received at the last step.

6. The last comment to make with reference to [5] is the strange reasoning whether the rewarding is to be part of the world or part of the device. The rewarding has been confidently treated as part of the world at the beginning of [5] and the reasoning about its belonging at the end of the paper is rather

odd. The authors answer this question themselves saying that if the students are allowed to mark their knowledge themselves they will all have excellent results.

Nevertheless, the confusion of Shane Legg and Marcus Hutter about the belonging of the rewards is reasonable. They ask themselves the question, if the rewards, in the case of human, is the feeling of pain and pleasure (with food, sex, music and other pleasure sources). The answer we would give is that human has not built-in reward. The success with human is evaluated by the world through the evolution. According to the evolution meaning, these who survive in the world are those who manage to survive and to reproduce. Human does not receive the evolution meaning by default. Actually, humans do never receive it. If one has lived in compliance with the evaluation meaning, that he is not aware about, then he will survive long enough and will be inherited in the next generation. This is the reason why people are looking for the meaning of life all their lives (they are looking for it, because they do not know it). As long as the feeling of pain and pleasure are concerned, this is not the meaning of life but an instinct. Thus, humans are born with some knowledge. They instinctively know that pain is bad, whereas pleasure is good. These instincts should not be trusted blindly. For example, the feeling of pain when the dentist pulls a tooth is misinform (as far as the dentist pulls the right tooth, of course). The bitter taste in food is instinctively perceived as bad but with time people get to like the taste of the coffee and the beer, for example. The feeling of pleasure is also often misinform.

Despite all the remarks we have made, the papers of Shane Legg and Marcus Hutter are very valuable to us, because they are the first to acknowledge the definitions in [1, 2]. Furthermore, the analysis of the mistakes made by our Swiss colleagues is also useful because it indicates what has not been explained sufficiently well. It is obvious that Shane Legg and Marcus Hutter have done serious work on this subject and it was useful to us to study their experience and the attempt to improve our definition. Where they have not understood the definitions of [1, 2], this is our fault, as we seem to have poorly explained, before. This paper was prepared on the basis of this analysis and we hope that it will help them and other researchers working in the field of Artificial Intelligence.

Acknowledgements

I want to thank professor Dimiter Skordev and professor Tinko Tinchev for their comments and advice to me regarding this paper. Furthermore, I would

like to thank profession Skordev that he challenged me at a time to write the definition [2] saying the following regarding definition [1]: 'This definition is not formal, at all. You claim that it can be formalized but what is not clear is how it can be done!' This was a constructive criticism and its result was definition [2].

References

.

[1] D o b r e v D. (2000) *AI – What is this*, In: PC Magazine – Bulgaria, November'2000, pp.12-13 (www.dobrev.com/AI/definition.html).

[2] D o b r e v D. (2005) *Formal Definition of Artificial Intelligence*, In: International Journal "Information Theories & Applications", vol.12, Number 3, 2005, pp.277-285 (www.dobrev.com/AI/).

[3] D o b r e v D. (2001) *AI – How does it cope in an arbitrary world*, In: PC Magazine – Bulgaria, February'2001, pp.12-13 (www.dobrev.com/AI/).

[4] L e g g S. and H u t t e r M. (2005) *A Universal Measure of Intelligence for Artificial Agents*, In: Proc. 21st International Joint Conf. on Artificial Intelligence (IJCAI-2005), pages 1509-1510, Edinburgh, 2005.

[5] L e g g S. and H u t t e r M. (2006) *A formal measure of machine intelligence*, In Proc. 15th Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn'06), pages 73-80, Ghent, 2006.

[6] L e g g S. and H u t t e r M. (2007) *Universal Intelligence: A Definition of Machine Intelligence*, Minds & Machines, 17:4 (2007) pages 391-444.